Solution Overview

# Cisco HPC Network Solutions for Microsoft Windows Compute Cluster Server 2003

**Cisco® InfiniBand and Ethernet Network Fabric Solutions for Microsoft Windows Compute Cluster Server (CCS) 2003**

## CHALLENGE

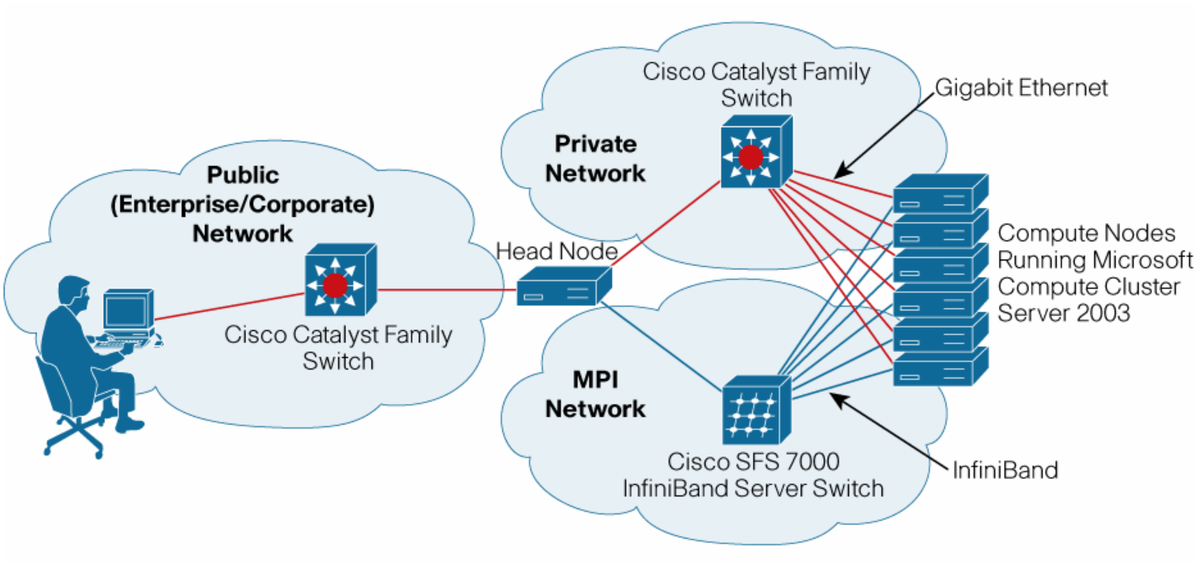
High-performance computing (HPC) clusters of industry-standard servers have emerged in recent years as the preferred method for implementing a supercomputer for computationally intensive tasks. Enterprises, research institutions, and governments use HPC clusters for a variety of purposes ranging from financial risk analysis to computational fluid dynamics, to weather and climate modeling, to analyzing underground oil and gas reservoirs. The applications and uses of HPC clusters are quite varied. The requirements of the underlying network are also quite varied. Customers are faced with a number of challenges. Among these are:

- What network interconnect offers the best performance or best value for my Microsoft Windows CCS application?
- What network solutions, products, and topologies are validated for Microsoft Windows CCS?

## SOLUTION

Cisco Systems® uniquely manufactures and supports multiple interconnects for HPC clusters, providing enterprises with a superior level of flexibility in terms of price, latency, and performance. A Gigabit Ethernet interconnect using a Cisco Catalyst® switch is quite suitable for loosely coupled, highly parallelized and parametric applications (Figure 1). Tightly coupled applications with high inter-nodal traffic rates are typically bandwidth- and latency-sensitive. These applications will benefit from the low latency, high bandwidth, and native Remote Direct Memory Access (RDMA) capabilities of InfiniBand using the Cisco SFS 7000 Series InfiniBand Server Switches. Customers can be assured of an appropriate interconnect solution for their HPC cluster application requirements.

**Figure 1.** Example HPC Cluster Based on Microsoft CCS and End-to-End Cisco Ethernet and InfiniBand Interconnects


Cisco Catalyst Family Switch
Gigabit Ethernet
Private Network
Public (Enterprise/Corporate) Network
Head Node
Compute Nodes Running Microsoft Compute Cluster Server 2003
Cisco Catalyst Family Switch
MPI Network
Cisco SFS 7000 InfiniBand Server Switch
InfiniBand

## NETWORK INTERCONNECT ARCHITECTURE FOR MICROSOFT WINDOWS CCS

Microsoft CCS offers five different cluster topologies (Table 1). The user is asked to select one of these topologies during the CCS installation process. CCS differentiates between two node types with each node supporting between one and three network interfaces:

- **Head node**—This node is the master of the cluster and may additionally support Remote Installation Services (RIS) for installing CCS images on compute nodes, and Internet Connection Sharing (ICS) Network Address Translation (NAT). One head node is used per cluster. The head node may also function as a compute node. The head node is also the DHCP server for the Private and the MPI network. The head node may or may not be the active directory and domain controller. However an active directory infrastructure and domain is required by Microsoft CCS.
- **Compute nodes**—These nodes perform the computational tasks and communicate with each other and the head node over the network interconnects. These nodes are a part of the same domain.

Three network types are used. A cluster may use one, two, or all three networks:

- **Public network**—The pre-existing enterprise or corporate network used to access the cluster. It is not necessarily a publicly accessible network, but a popular term to differentiate it from the private and Message Passing Interface (MPI) networks. The head node is always connected to the public network. Compute nodes may be connected to the public network for management purposes. A public interface is required for the head node but not required for the compute nodes.
- **Private network**—A dedicated cluster network interconnecting the head node and compute nodes. It carries management and deployment traffic. It also carries MPI traffic if a dedicated MPI network is not used. This private management network also performs job scheduling and initialization, Network file sharing, and remote event logging.
- **MPI network**—A dedicated high-speed network for inter-nodal message passing traffic. This network, if used, is most appropriate for the low-latency and high-bandwidth capabilities of InfiniBand using Cisco SFS 7000 Series switches. An IP address has to be assigned to the MPI Network. If using InfiniBand as the interconnect for the MPI Network, IPoIB drivers must be installed. Lower latency and higher bandwidth can be realized using the WSD (Winsock Direct) provider for InfiniBand.

### Supported Windows CCS Network Topologies

Table 1 lists the five supported Microsoft Windows CCS topologies with suggested network interconnects.

**Table 1.** Windows CCS Topologies

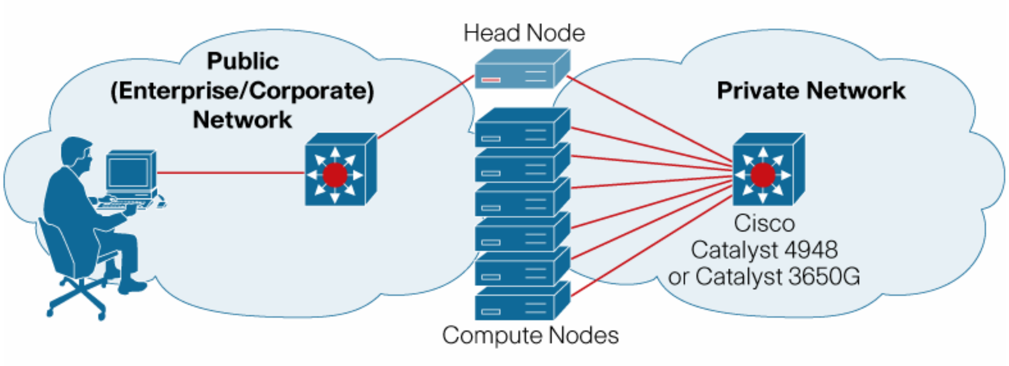
| Topology | Description | Network Performance Requirement | Network Interconnect Technology Guideline | | |
|---|---|---|---|---|---|
| | | | **MPI Network** | **Private Network** | **Public Network** |
| **1** | Public network connection to head node only; dedicated private network; NAT used | High | Not used | Gigabit Ethernet | Ethernet (10/100/1000) |
| **2** | Public network connection to all nodes; dedicated private network; NAT optional | High | Not used | Gigabit Ethernet or InfiniBand | Ethernet (10/100/1000) |
| **3** | Public network connection to head node only; dedicated private network to all nodes; dedicated MPI network to all nodes; NAT used | High | Gigabit Ethernet | Gigabit Ethernet | Ethernet (10/100/1000) |
| | | Very high | InfiniBand | | |
| **4** | Public network connection to all nodes; dedicated private network to all nodes; dedicated MPI network to all nodes; NAT optional | High | Gigabit Ethernet | Gigabit Ethernet | Ethernet (10/100/1000) |
| | | Very high | InfiniBand | | |
| **5** | Public network connection to all nodes | Low to medium | Not used | Not used | Ethernet (10/100/1000) |

### Topology Example 1: Dedicated Private Network; Public Network to Head Node Only

Figure 2 shows a sample topology. This is a simple, high-performance solution using Gigabit Ethernet for the private network. This network would handle management, image loading, and computational traffic between the nodes.

The Cisco Catalyst 3560G-24TS Switch with 32 Gbps of forwarding bandwidth is an appropriate choice for smaller clusters with up to 24 nodes. Clusters with up to 48 nodes and those requiring full bisectional bandwidth and low latency can use the Catalyst 4948 Switch with 96-Gbps forwarding bandwidth and 5-microsecond (usec) latency. Larger clusters can use a modular switch such as the Catalyst 6500 Series Switch or a combination of top-of-rack and aggregation switches to achieve the required density. The Head Node provides DHCP and RIS Services on the private interface.

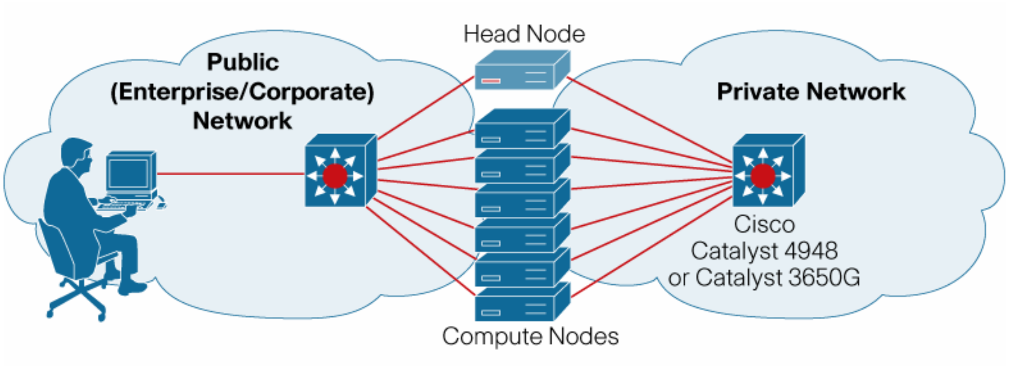
**Figure 2.** Dedicated Private Network for Compute Nodes and Public Network Connection to Head Node Only



### Topology 2: Dedicated Private Network; Public Network to all Nodes

Topology 2 (Figure 3) does not rely on or require the private network for management connectivity of the compute nodes. All nodes including the head node are connected to the public network; therefore ICS NAT is not required on the head node for public-to-private network connectivity. Switching product choices for the private network are the same as for Topology 1.

**Figure 3.** All Nodes Connected to Public and Private Networks



### Topology 3: Dedicated MPI Network; Public Network Connection to Head Node

A dedicated MPI network offers the best performance of any network scenario because only MPI messages traverse the MPI network. Applications with high inter-nodal traffic or high bandwidth requirements can benefit from this topology, particularly when a low-latency, high-bandwidth InfiniBand MPI network is used. The Cisco SFS 7000 InfiniBand Server Switch features full non-blocking 10-Gbps performance per node for up to 24 nodes in a 1 rack unit (RU) form factor. Larger clusters can use the modular Cisco SFS 7008 (up to 96 ports); Cisco SFS 7012P (144 ports); or Cisco SFS 7024P (288 ports) in a single tier or top-of-rack and aggregation arrangement.

Two implementations are shown for this topology: one based on Ethernet for the MPI network, and another higher-performance implementation based on Cisco SFS 7000 InfiniBand (Figures 4 and 5).

**Figure 4.** Dedicated Private and MPI Networks Using Ethernet

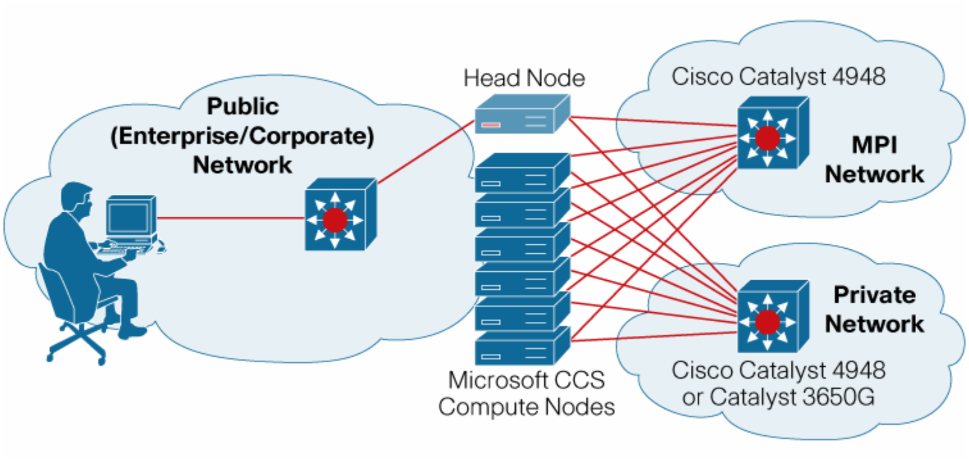
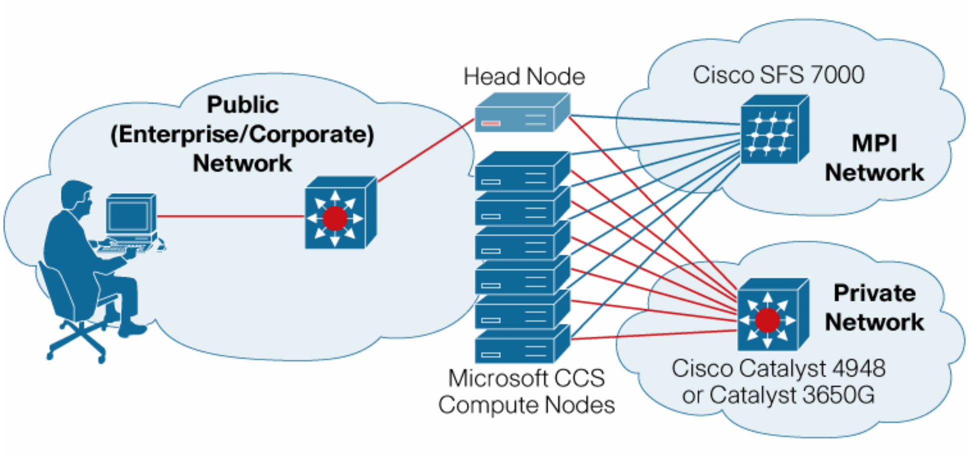


**Figure 5.** Dedicated Private Network Using Ethernet and MPI Network Using Cisco SFS 7000 InfiniBand



### Topology 4: Dedicated MPI Network with all Nodes Connected to Public Network

Topology 4 also uses a dedicated MPI network, so offers similar performance benefits to Topology 3. But like Topology 2, all nodes are connected to the public network and so negate the need for ICS NAT in the head node. Figures 6 and 7 show two implementations of Topology 4: one based on an Ethernet MPI network, the other on an InfiniBand MPI network.

**Figure 6.** All Nodes Connected to all the Network over Ethernet

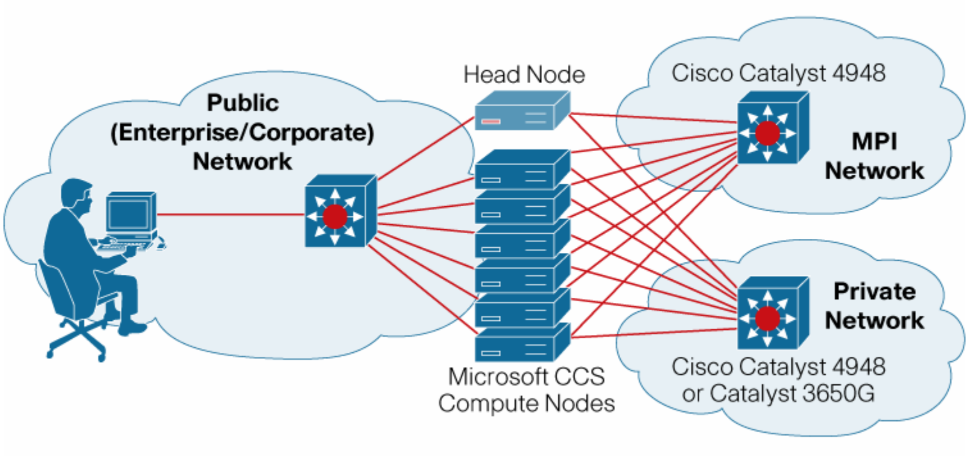
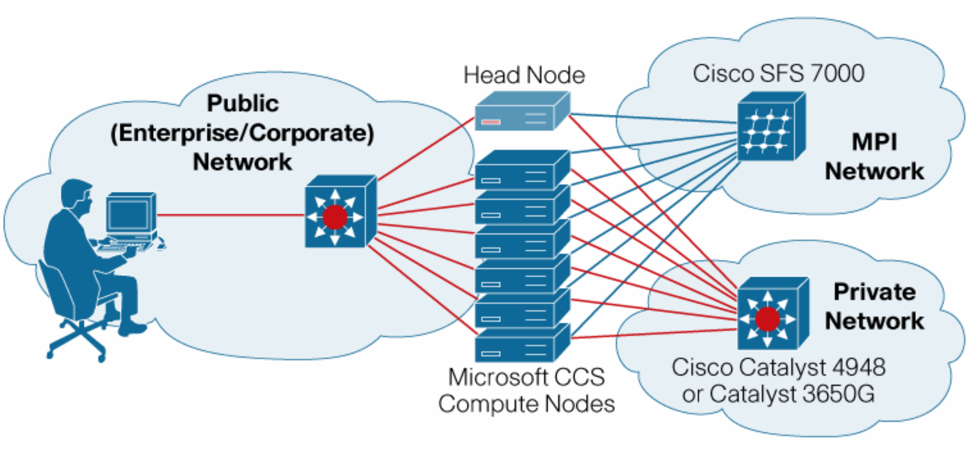

Public (Enterprise/Corporate) Network
Head Node
Microsoft CCS Compute Nodes
Cisco Catalyst 4948
MPI Network
Private Network
Cisco Catalyst 4948 or Catalyst 3650G

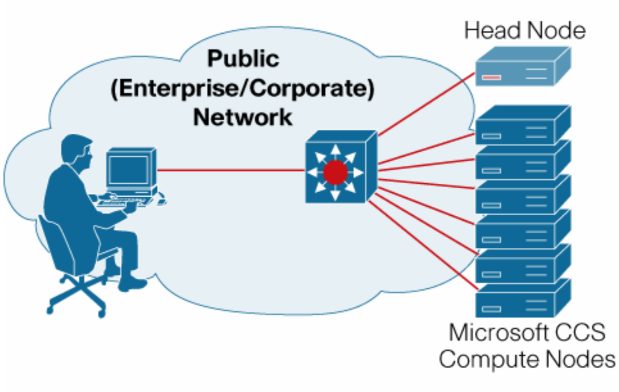
**Figure 7.** Dedicated MPI Networks Using Cisco SFS 7000 InfiniBand


Public (Enterprise/Corporate) Network
Head Node
Microsoft CCS Compute Nodes
Cisco SFS 7000
MPI Network
Private Network
Cisco Catalyst 4948 or Catalyst 3650G

### Topology 5: Single Public Network Connection to all Nodes

Topology 5 (Figure 8) is the simplest of all supported Windows CCS topologies with all nodes connected to a single public network. All management and MPI traffic will traverse the public network switch or switches, so performance will very much depend on the capacity and switching capabilities of the public (corporate) network infrastructure.

**Figure 8.** Simple Single Public Network Connection for all Nodes



### Topology Variations for Larger Clusters

Depending on server form factor (blade server, 1-RU server), a typical 42-RU equipment rack will normally accommodate between 32 and 40 servers. In these configurations, a single switch per network is all that is required (for example, a Cisco Catalyst 4948 with 48 Gigabit Ethernet interfaces for a private network).

Larger clusters of 64 nodes, 128 nodes, or many hundreds of nodes require more involved network topologies spanning multiple equipment racks. With any HPC cluster topology involving multiple switches, it is important to consider blocking and network oversubscription factors in order to achieve expected performance.

Figure 9 shows two multiple-rack Ethernet designs. The first design shows two racks interconnected by a 20-Gbps Cisco EtherChannel® connection comprising two 10 Gigabit Ethernet links between two Catalyst 4948-10GE Switches deployed at the top of the rack. This design could apply to a 64-node Windows CCS cluster using Topology 1 or 2. The second design shows an even larger cluster using four Catalyst 4948-10GE Switches at the top of the rack, but aggregating to a Catalyst 6500 Series Switch. The design as shown scales to a 512-node cluster (based on 32 compute nodes per rack) with an oversubscription factor of only 1.6:1.

Where a dedicated InfiniBand-based MPI network is deployed (Topology 3 and 4), a low oversubscription factor on the Ethernet network is less important. In this case, single or EtherChannel Gigabit Ethernet links could be substituted for the 10 Gigabit Ethernet links aggregating to Catalyst 6500 Series or Catalyst 4948 Switches where density permits.

**Figure 9.** Example Large Ethernet HPC Cluster Network Design Using Cisco Catalyst Switches

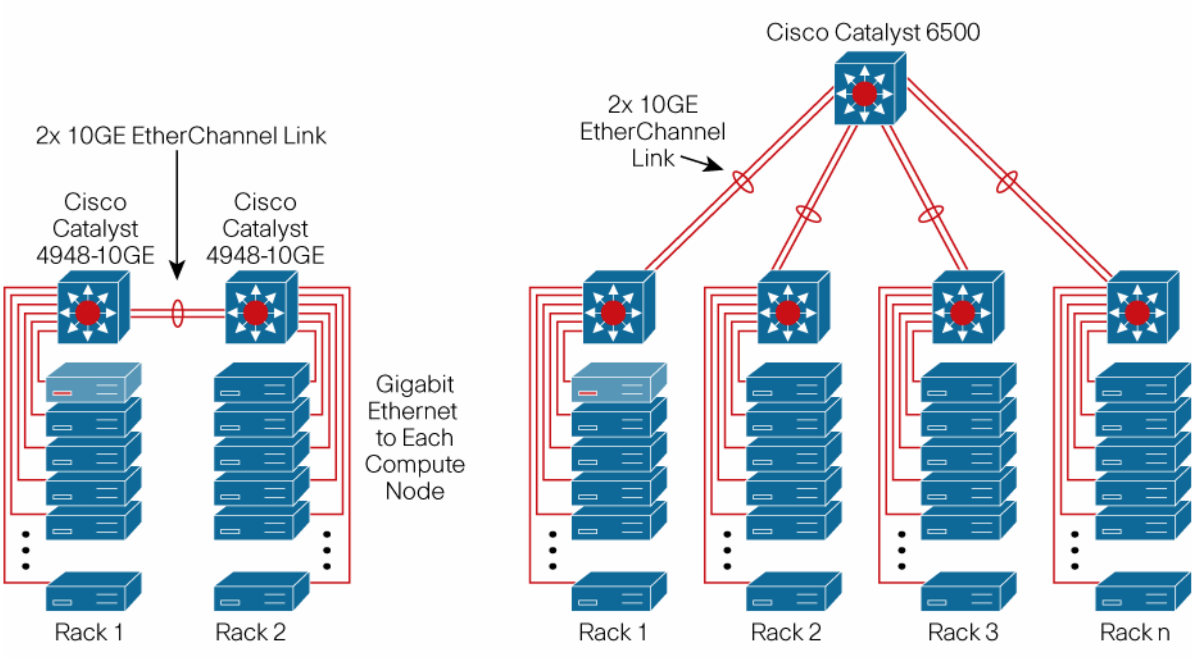
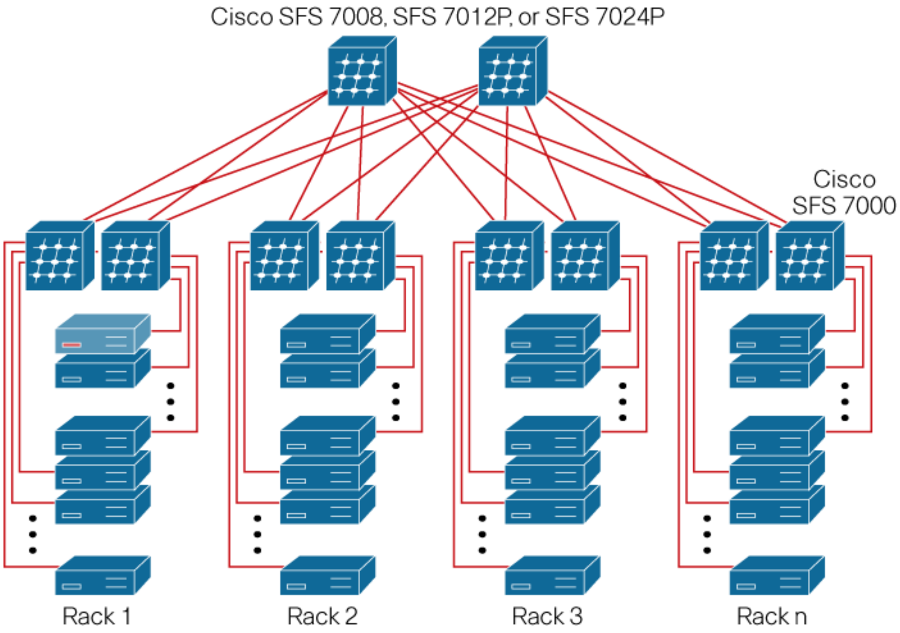


Figure 10 shows a larger multi-rack InfiniBand MPI network deployment based on the Cisco SFS 7000 Series InfiniBand Server Switches. In this instance, two 24-port Cisco SFS 7000 Series switches connect to each of the compute nodes through 10-Gbps InfiniBand links. Remaining InfiniBand links are aggregated to a modular Cisco SFS 7000 Series switch (Cisco SFS 7008, SFS 7012P, or SFS 7024P).

**Figure 10.** Large InfiniBand HPC Cluster Network Using Cisco SFS 7000 Series InfiniBand Server Switches (for Private and/or MPI Networks)



### Choosing a Topology

Application requirements and characteristics determine network topology requirements. Tightly coupled applications with high inter-nodal traffic rates (such as Computational Fluid Dynamics [CFD] applications) will see the most benefits from a dedicated, low-latency, high-bandwidth MPI network based on Cisco SFS 7000 InfiniBand Server Switches (shown in Topology 3 and 4). More parallelized or loosely coupled applications may perform adequately with a private Gigabit Ethernet network (Topology 1 and 2).

## THE CISCO HPC CLUSTER PRODUCT AND SOLUTION SUITE

Cisco manufactures a broad range of Ethernet, InfiniBand, and Fibre Channel switching products. Table 2 outlines a selection of those products shown in the previous topologies and applicable to HPC clustering solutions using Microsoft Windows Compute Cluster Server.

**Table 2.** Cisco Catalyst Switches for HPC Clustering with Windows CCS

| Switch | Technology | Capabilities and Application |
|---|---|---|
| **Catalyst 3560G-24TS** | 1-RU, 24-port Gigabit Ethernet (GE) switch (10/100/1000) + 4 Small Form-Factor Pluggable (SFP) interfaces | 32 Gbps forwarding bandwidth Suitable for clusters up to 24 nodes |
| **Catalyst 4948** | 1-RU, 48-port GE switch | 96-Gbps forwarding bandwidth, non-blocking, low latency (approximately 5 usec at wire speed). High-performance GE switch suitable for clusters up to 48 nodes or at top-of-rack |
| **Catalyst 4948-10GE** | 1-RU, 48-port GE switch with two 10 GE interfaces (X2 pluggable) | 136-Gbps forwarding bandwidth, non-blocking, low latency (approximately 5 usec at wire speed). High-performance GE switch suitable for top-of-rack deployment in larger Ethernet-only clusters. |
| **Catalyst 6504-E** | 5-RU, 4-slot; 144 GE; 288 x 10/100; 12 x 10 GE | High-density, high-availability, high-performance architecture featuring redundant supervisors, fans, and power supplies with modular operating system and integrated management tools |
| **Catalyst 6506-E** | 12-RU, 6-slot; 240 GE; 480 x 10/100; 20 x 10 GE | High-density, high-availability, high-performance architecture featuring redundant supervisors, fans, and power supplies with modular operating system and integrated management tools. |
| **Catalyst 6509-E** | 15-RU, 9-slot; 384 GE; 768 x 10/100; 32 x 10 GE | High-density, high-availability, high-performance architecture featuring redundant supervisors, fans, and power supplies with modular operating system and integrated management tools. |
| **Catalyst 6509-NEB-A** | 21-RU, 9-slot; 384 GE; 768 x 10/100; 32 x 10 GE | High-density, high-availability, high-performance architecture featuring redundant supervisors, fans, and power supplies with modular operating system and integrated management tools. |
| **SFS 7000P** | 1-RU, 24-port 4X InfiniBand switch | Fully non-blocking, top-of-rack switch supports optical and CX4 interfaces. Includes embedded subnet manager. |
| **SFS 7008P** | 6-RU, modular InfiniBand switch (up to 96 4X ports or 24 12X ports) | Fully non-blocking switch supports optical and CX4 interfaces. Suitable as single multi-rack switch or aggregation switch in larger clusters. Includes embedded subnet manager. |
| **SFS 7012P** | 7-RU, modular InfiniBand switch (up to 144 4X ports) | Fully non-blocking InfiniBand switch supports optical and CX4 interfaces. Suitable as single multi-rack switch or aggregation switch in larger clusters. |
| **SFS 7024P** | 14-RU, modular InfiniBand switch (up to 288 4X ports) | Fully non-blocking InfiniBand switch supports optical and CX4 interfaces. Suitable as single multi-rack switch or aggregation switch in larger clusters |

## WHY CISCO

- **Multiple fabric technologies**—Cisco is currently the only vendor to offer a variety of standards-based fabric technologies—Gigabit Ethernet, 10 Gigabit Ethernet, InfiniBand, and Fibre Channel. Customers can choose the most appropriate solution and mix of fabrics for their application.
- **Any application**—Cisco supports and has validated solutions with a variety of independent software vendor (ISV) applications.
- **End-to-end fabric management**—Only Cisco can offer an end-to-end management solution over multiple fabric technologies for greater availability and serviceability.
- **Unparalleled service and support**—All fabrics are supported by Cisco's renowned worldwide service and support.

## FOR MORE INFORMATION

For more information about Cisco HPC solutions, visit www.cisco.com/go/hpc or contact your local account representative.

CISCO SYSTEMS

**Corporate Headquarters**
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 526-4100

**European Headquarters**
Cisco Systems International BV
Haarlerbergpark
Haarlerbergweg 13-19
1101 CH Amsterdam
The Netherlands
www-europe.cisco.com
Tel: 31 0 20 357 1000
Fax: 31 0 20 357 1100

**Americas Headquarters**
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
www.cisco.com
Tel: 408 526-7660
Fax: 408 527-0883

**Asia Pacific Headquarters**
Cisco Systems, Inc.
168 Robinson Road
#28-01 Capital Tower
Singapore 068912
www.cisco.com
Tel: +65 6317 7777
Fax: +65 6317 7799

**Cisco Systems has more than 200 offices in the following countries and regions. Addresses, phone numbers, and fax numbers are listed on the Cisco.com Website at www.cisco.com/go/offices.**

Argentina • Australia • Austria • Belgium • Brazil • Bulgaria • Canada • Chile • China PRC • Colombia • Costa Rica • Croatia • Cyprus • Czech Republic Denmark • Dubai, UAE • Finland • France • Germany • Greece • Hong Kong SAR • Hungary • India • Indonesia • Ireland • Israel • Italy Japan • Korea • Luxembourg • Malaysia • Mexico • The Netherlands • New Zealand • Norway • Peru • Philippines • Poland • Portugal Puerto Rico • Romania • Russia • Saudi Arabia • Scotland • Singapore • Slovakia • Slovenia • South Africa • Spain • Sweden Switzerland • Taiwan • Thailand • Turkey • Ukraine • United Kingdom • United States • Venezuela • Vietnam • Zimbabwe

Printed in USA C11-354223-00 06/06